



## Lettres, mots, textes - Clefs d'accès à l'écrit numérique

Bénédicte Pincemin

### ► To cite this version:

Bénédicte Pincemin. Lettres, mots, textes - Clefs d'accès à l'écrit numérique. Journée scientifique "Sensibilisation aux outils informatiques et statistiques d'aide à l'analyse des textes", Feb 2001, Reims, France. pp.59-87. halshs-00168992

**HAL Id: halshs-00168992**

**<https://shs.hal.science/halshs-00168992>**

Submitted on 21 Apr 2009

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

**Bénédicte PINCEMIN**  
CNRS & LLI, université Paris 13

## Lettres, mots, textes Clefs d'accès à l'écrit numérique

### 1. Du bon usage des moteurs de recherche sur le texte intégral

#### 1.1. L'homme de Lettres, la littérature et l'ordinateur

Parmi les ressources apportées aujourd'hui par l'informatique à l'homme de Lettres, un nouveau mode d'exploration du texte intégral d'œuvres se dessine. Outil pour une lecture systématique selon des critères explicités, embrassant les textes dans leur ensemble, l'ordinateur se révèle surtout par l'ampleur quantitative des corpus qui peuvent lui être soumis : telle recherche sur le vocabulaire, inenvisageable à l'échelle d'une vie humaine, devient accessible en quelques secondes, puis révisable, reformulable.

A vrai dire, les craintes ou les engouements abusifs pour l'usage de l'informatique en littérature viennent en bonne part d'une méprise sur le rôle de l'ordinateur dans la lecture : sa capacité d'enregistrement et sa puissance de calcul n'ont pas le même empan que la mémoire et le raisonnement humain, mais surtout elles ne sont pas de même nature. L'ordinateur n'est pas un interlocuteur, mais un support de travail, un outil : toute l'activité de lecture et d'interprétation, l'intelligence du texte, et sa perception même comme une expression esthétique ou cognitive, tout ceci reste entièrement entre les mains de l'homme.

Quant à la pertinence de procéder à des calculs et d'appliquer des modèles statistiques aux textes, les résultats de la lexicométrie illustrent la puissance de suggestion et la mise en valeur de caractéristiques inaperçues et significatives, pourvu que l'interprète se garde tant de la fascination que de l'effroi vis à vis des formules mathématiques sous-jacentes (Muller 1985).

#### 1.2. Les traitements disponibles

Des logiciels spécialisés dans l'analyse de textes sont de riches boîtes à outils pour le professionnel de la linguistique ou de la littérature. De même que disposer d'une scie ne dispense pas, ni n'impose, d'avoir une clé à molette, diverses perspectives sur le texte sont outillées : par exemple morphosyntaxique (Traitements Automatiques des Langues, avec un analyseur comme INTEx), ou documentaire (consultation, localisation et inventaire de passages et d'occurrences, que fournit typiquement un concordancier comme X-COR), ou encore statistique et visuelle (Hyperbase est un des outils les plus perfectionnés dans ce domaine).

Et pourtant, l'outil courant livré avec le texte électronique — si tant est qu'il y en ait un — reste comme en deçà de tous ces axes d'investigation. Les textes littéraires électroniques à large diffusion

sont mis à disposition via des « sites bibliothèques » sur internet<sup>1</sup>, ou édités et commercialisés sous forme de cédéroms d'œuvres complètes<sup>2</sup>. Un hypertexte minimal<sup>3</sup> permet de feuilleter le texte, mais ne suffit guère à justifier le support électronique : l'écran n'offre pas du tout la même qualité de lecture que le livre traditionnel<sup>4</sup>. La fonctionnalité clef d'accès au texte est alors le moteur de recherche : repérage de toutes les occurrences d'un mot ou d'une expression, et affichage des passages correspondants.

### 1.3. Un point de vue linguistique sur les moteurs de recherche

Quatre principales raisons conduisent donc à centrer cet exposé sur les apports possibles des moteurs de recherche pour le travail sur un corpus de textes. (i) La possibilité d'interroger une collection de textes électroniques par un mot ou une équation booléenne est la principale interface disponible sur les ressources largement diffusées sur internet ou sur les textes édités sur cédérom. (ii) L'examen détaillé de cet outil standard est complémentaire de la présentation de logiciels spécialisés, tels que ceux présentés dans les exposés précédents. (iii) Par contraste avec d'autres interfaces standard, les moteurs de recherche travaillent au niveau de la matière même du texte, non d'une description surajoutée (« méta »-textuelle) qui fixe et ferme les points d'entrée et les parcours (mots-clés, répartition dans des dossiers thématiques). Leur fonctionnement pose des questions intéressantes au plan linguistique, et par principe laisse une grande liberté de perspectives à l'étude littéraire. (iv) Mon expérience dans la conception d'un outil de calcul de similarités textuelles dans le cadre d'une application documentaire<sup>5</sup> m'a conduit à examiner

---

<sup>1</sup> Association des Bibliophiles Universels :

<http://cedric.cnam.fr/ABU/>

Le fonds numérisé de la Bibliothèque Nationale de France :

<http://gallica.bnf.fr/>

Biblionet : <http://minotaure.bibliopolis.fr:7999/>

pour ne citer que quelques grands sites librement accessibles, et offrant non seulement une version électronique des textes (à afficher ou télécharger) mais aussi une interface de consultation et d'interrogation.

<sup>2</sup> Les réalisations d'Acamédia (Chateaubriand, Alexandre Dumas) ou du Catalogue des Lettres (Flaubert, Zola) sont de bons exemples, à distinguer d'autres cédéroms littéraires ne faisant qu'une place secondaire au texte de l'œuvre pour privilégier son entour, et en développant l'aspect multimédia ou documentaire.

<sup>3</sup> Il est difficile de faire moins que de poser des liens hypertextes donnant accès aux différents livres, puis aux pages successives et aux sections qui le divisent (chapitres, actes et scènes, etc.). L'hypertexte ne prend toute son ampleur que lorsqu'il propose des parcours échappant aux repérages déjà présents dans les éditions traditionnelles (table des matières, index, notes...), par exemple lorsque tout mot du texte, cliquable, devient un relais pertinent au travers d'une multiplicité de vues possibles sur le texte (cf. la notion de « texte dynamique », et l'exemple d'Hyperbase).

<sup>4</sup> Le bureau sans papier relève de l'utopie, il n'y a d'ailleurs jamais autant eu de papier consommé et accumulé qu'à l'ère du numérique. Penser la lecture électronique comme se substituant à la lecture papier est démenti par les multiples « infériorités » du texte affiché à l'écran : luminosité fatigante de l'écran ; encombrement, fragilité, dépendance énergétique de la machine ; perte de repères matériels comme l'épaisseur du livre, le grain du papier, et dont l'importance cognitive est avérée ; caducité rapide des formats informatiques, volatilité des pages internet....

<sup>5</sup> DECID : serveur de Diffusion Electronique Ciblée d'Informations et de Documents sur l'intranet EDF, cf. (Bommier-Pincemin 1999).

l'état de l'art. La connaissance des aspects techniques (comment ça marche) est un préalable à la compréhension de leurs forces et de leurs faiblesses (pourquoi ça (ne) marche (pas)), compréhension ensuite mise à profit dans un savoir-faire (comment utiliser au mieux, paramétrer), dans une évaluation critique des offres (qu'est-ce qui est vraiment utile dans les « plus » de tel ou tel outil), et dans la conception de nouvelles formes d'accès sémantique aux textes.

#### **1.4. Logique de l'exposé**

C'est un tel parcours auquel je vous invite. L'étude des moteurs de recherche se déroulera en trois temps : une introduction aux modèles et aux techniques sous-jacents ; la signification linguistique des opérateurs booléens, suite à leur transposition au texte intégral ; l'exploitation des résultats d'une interrogation, pour leur interprétation juste.

Ensuite, en prenant du recul, il ressort trois manières de considérer les textes au cours des traitements automatiques, trois clés d'accès comme l'annonce le titre : les lettres, point de vue graphique permis par l'écriture alphabétique ; les mots, unités linguistiques, qui font appel à une connaissance sur la langue ; la textualité, qui mobilise la notion de contexte et fonde une sémantique différentielle comme celle de François Rastier (1987).

L'objectif ici est pragmatique : en tant qu'utilisateurs, il nous intéresse moins de juger l'outil, « bon » ou « mauvais », que de prendre du recul par rapport à son mode de fonctionnement pour en exploiter au mieux les possibilités. Les calculs et l'automatisation introduits par l'informatique ne réduisent pas le lecteur à la passivité, ils ne font pas l'analyse littéraire : bien au contraire, l'explicitation des hypothèses à soumettre à l'analyse, l'art et la manière d'employer les outils pour étayer ou éprouver ces hypothèses, sont de nouveaux lieux d'implication critique et interprétative.

## **2. Principes techniques de fonctionnement des moteurs de recherche**

### **2.1. L'indexation**

Puisque le moteur vise à retrouver des éléments dans les textes, cette première étape, l'indexation, consiste à analyser les textes pour repérer les éléments susceptibles d'être tout ou partie de l'objet d'une recherche.

L'articulation privilégiée est bien entendu celle des mots. Une approximation courante consiste à se contenter d'un découpage : un mot est une chaîne de caractères alphabétiques délimitée par des espaces ou des ponctuations. Cette approximation est la première pierre d'achoppement de l'utilisateur néophyte, puisque l'unité graphique ne se superpose pas avec l'unité linguistique : en particulier, la description d'un mot dans un dictionnaire englobe toutes ses variations de conjugaison et d'accord. L'unité graphique ne correspond pas non plus à une unité de sens, à la manière dont les mots clés en documentation sont associés conventionnellement à des thématiques, via un thesaurus.

L'indexation consiste alors à produire un « fichier inverse », qui à chaque mot associe l'ensemble de ses localisations dans les textes.

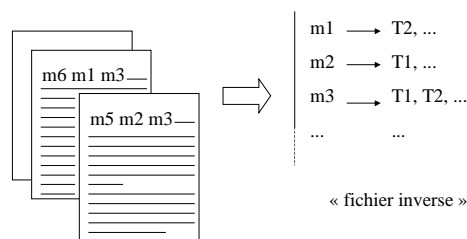


Figure 1 : l'indexation (au sens informatique)

Autrement dit, disposant pour chaque texte de l'information sur les mots qu'il contient, on souhaite avoir pour chaque mot l'ensemble des textes qui le contiennent. C'est une sorte d'index, qui permet d'accélérer les recherches (pour chercher un mot donné, au lieu de tout relire, l'ordinateur a les indications de ses localisations pour le trouver directement). Cette manière de faire ménage aussi la possibilité d'introduire quelques modifications entre l'ensemble des chaînes de caractères telles qu'elles figurent dans le texte, et les chaînes de caractères que l'on peut vouloir associer au texte (sélection, extension,...).

Ainsi, pour l'efficacité globale du système, tous les mots ne sont pas conservés comme éléments possible de recherche. La plupart des moteurs n'enregistrent pas les « mots vides », à savoir les mots grammaticaux tels que prépositions, déterminants, conjonctions : le présupposé est que les recherches sont d'ordre thématique et portent sur les mots lexicaux. Cela s'avère limitant pour explorer certains aspects stylistiques ou énonciatifs. Mais, même au plan thématique et sémantique, la focalisation sur le lexique ne permet pas de prendre en compte le rôle d'autres unités linguistiques, comme les ponctuations par exemple (Bourion 1998).

Les moteurs prennent généralement en compte les hapax (mots qui n'apparaissent qu'une seule fois), quelquefois écartés des analyses statistiques. En effet, les hapax peuvent représenter de l'ordre de la moitié du vocabulaire d'un texte, et concentrent les points anecdotiques, les erreurs de transcription, etc.

Une pondération peut intervenir pour moduler l'importance que peut prendre un mot lors de l'indexation (choix des mots les plus caractéristiques du texte) ou de la recherche (contribution à un classement par ordre de pertinence des textes retrouvés). Les formules de calcul de pondération combinent traditionnellement une mesure de la présence et du rôle central du mot dans le texte (souvent basée sur sa fréquence) et une mesure de la rareté du mot dans le corpus (si le mot est employé dans presque tous les textes, il n'est pas aussi significatif que s'il est propre à quelques textes).

Certains moteurs de recherche ont proposé que l'utilisateur puisse indiquer des pondérations pour les mots de sa requête, l'idée étant qu'il exprime ainsi ce qui doit rester au cœur de la recherche, et ce qui est secondaire. Cette fonctionnalité s'avère décevante, car l'utilisateur ne dispose d'aucun moyen pour ajuster les valeurs numériques et percevoir la réelle influence de ses choix. Une voie plus fructueuse serait la possibilité de qualifier les différents termes de recherche (au lieu de les quantifier), selon qu'ils contribuent significativement au sujet de la recherche, qu'ils sont des critères de sélection fiable (parce que très particuliers, très précis, univoques), ou que leur absence est pénalisante ou non.

## 2.2. L'interrogation

Lancer une recherche sur un mot permet de localiser des occurrences. La consultation du contexte immédiat du mot est alors essentielle pour vérifier l'adéquation entre les attentes que l'on a voulu traduire dans la requête, et ce qui a été effectivement repéré. En particulier, il est dangereux de se contenter d'un nombre d'occurrences total (tel mot apparaît tant de fois dans l'œuvre d'Untel) sans examiner au moins un échantillon des contextes d'apparition. Revenir au texte est essentiel pour s'apercevoir d'emplois imprévus, comme pour prendre connaissance du champ sémantique de ce mot dans le corpus étudié (discerner la gamme de ses emplois).

Les deux modes de recherche principaux (le booléen et le vectoriel) se différencient dès lors que la recherche combine plusieurs mots.

L'interrogation booléenne utilise des opérateurs comme OU, ET, SAUF, et éventuellement des parenthèses<sup>6</sup>, pour exprimer les rapports entre les différents mots mentionnés, alors que la recherche basée sur le modèle de l'espace vectoriel se présente comme une libre énumération de mots<sup>7</sup>.

La force de l'expression booléenne tient à son caractère structuré (les interrelations entre les mots sont typées). Les critères sur lesquels s'effectue la sélection des textes présentés en réponse sont nettement explicités. Cela donne les moyens d'élaborer des stratégies d'interrogation, puisque, pour peu qu'il ait acquis un certain savoir-faire, l'utilisateur peut ajuster et reformuler sa requête en contrôlant les effets de ses retouches.

En revanche, par définition, le booléen est un modèle binaire, basé sur l'opposition vrai / faux, bref sur une logique du tout ou rien. L'équation booléenne fonctionne comme un filtre, et ne fait que séparer les textes qui répondent aux critères de la requête, de ceux qui n'y répondent pas. Il n'y a pas d'à peu près (tel texte correspondant presque à l'équation est écarté, et donc occulté), ni de nuances. D'où les difficultés pour l'exploitation du paquet de textes sélectionné : il n'y a pas d'ordre logique pour les parcourir, et le volume total des réponses oscille par à coups entre le prolixe (trop de OU...) et le lacunaire (trop de ET !).

Face à cette exploitation délicate et précise d'un formalisme rigoureux, le mode vectoriel séduit par sa simplicité d'expression : l'évocation du thème de recherche procède par la mention de quelques mots, sans mise en forme particulière. Bien entendu, les interrelations entre les mots n'en existent pas moins, et le modèle

---

<sup>6</sup> Les interfaces des moteurs de recherche grand public sur internet tendent à masquer ces complexes (?) délimitateurs structurants, au profit :

- soit d'une limitation très brutale des possibilités : choix d'un seul opérateur qui s'applique à tous les termes, via les boutons « au moins un mot » (= OU), « tous les mots » (= ET), « expression exacte » (ou l'anglicisme « phrase exacte ») (= adjacence).

- soit d'une présentation aberrante, sous forme d'un menu déroulant pour le choix de l'opérateur précédant chaque fenêtre d'introduction d'un mot-clé supplémentaire : cela présente les opérateurs OU, ET, PRES DE, ADJACENT, etc. comme associés à 1 mot (opérateurs unaires), alors qu'ils décrivent des dépendances entre plusieurs (opérateurs n-aires). Dès que l'on utilise des opérateurs différents dans la même requête, la signification de la requête est confuse, et gérée par des conventions implicites : priorités sur les opérateurs ? ou regroupements à gauche d'abord ? ou... ?

<sup>7</sup> C'est typiquement le premier mode d'utilisation des moteurs de recherche sur internet. Le *Modèle de l'Espace Vectoriel* a été mis au point dans l'équipe de Salton, cf. (Salton & McGill 1983).

vectoriel résout cette question de façon purement quantitative et cumulative. Concrètement, chaque mot indépendamment apporte une contribution au calcul du score de pertinence d'un texte vis à vis de la requête : contribution nulle (ou pénalisante<sup>8</sup>) s'il ne figure pas dans le texte, contribution positive s'il y figure, et d'autant plus forte qu'il joue un rôle important (évalué par une pondération, cf. § indexation). C'est une somme de contributions mot à mot qui détermine les textes trouvés. Or, par sa facture même, le score est leurré par (i) un mot isolé atypique, surévalué par sa pondération (somme réduite à un terme), ou inversement (ii) un assemblage sans cohérence de mots d'incidence mineure dispersés dans le texte (cumul de multiples petites valeurs).

Pour mention, il existe aussi l'interrogation dite « en langue naturelle », où l'expression du thème de recherche prend la forme d'une demande. Cela est souvent présenté comme une prouesse technique (« la machine vous comprend, dans votre langue de tous les jours ») et comme un supplément de convivialité. Mais à l'usage, on préfère entrer au clavier quelques mots en relation avec le thème de recherche, que de rédiger une demande. De plus, la difficulté principale n'est peut-être pas d'exprimer le thème de la recherche sous une forme recevable par le moteur de recherche, que de maîtriser la manière dont le moteur va interpréter la requête et procéder à la recherche, autrement dit comment bien lui faire comprendre ce que l'on cherche. Quant à l'analyse linguistique<sup>9</sup> de la demande formulée, d'une part elle est effectivement très complexe (tolérance aux erreurs, portée des négations, résolution des anaphores...) et donc met durement à l'épreuve les performances et la robustesse des analyseurs ; d'autre part si l'analyse consiste à transposer la demande sous la forme d'une équation booléenne élémentaire, alors il serait plus sûr, plus puissant et plus efficace de l'écrire directement.

### **3. Interprétation linguistique des opérateurs booléens**

#### **3.1. Des langages documentaires à la langue des textes**

Traditionnellement, les systèmes de recherche documentaire ont été développés pour l'interrogation de fonds catalogués et enregistrés dans une base de données. Chaque document y est décrit par des champs structurés (auteur, date de publication, titre...) dont le contenu suit certaines règles (par exemple, nombre limité de mots clés, pris dans une liste d'autorité). Il y a donc un langage de représentation des documents, auquel correspond le langage d'interrogation de la base.

Avec la prolifération des moyens d'édition et de publication, le volume des documents électroniques disponibles, ne serait-ce que sur internet, a conduit à la mise en place de moteurs de recherche, sur le modèle des systèmes documentaires classiques. Cette transposition ne se fait pas sans introduire de multiples décalages, pas toujours perçus : les mots extraits des textes intégraux ne fonctionnent pas comme des mots-clés, et leur combinaison dans les textes ne met pas en jeu les mêmes relations que celles des champs d'une base bibliographique.

---

<sup>8</sup> Cela dépend de la formule utilisée pour le calcul de pertinence, et s'applique indifféremment à tous les mots manquants.

<sup>9</sup> Y compris la prise en compte du « genre textuel » dont relèvent ces requêtes : brièveté, formules de demande (« je cherche des informations sur... »), fréquence des coquilles, etc.

Pour utiliser au mieux les formalismes disponibles, il s'agit de comprendre quel phénomène linguistique (à l'œuvre dans le texte intégral) peut être associé à tel ou tel opérateur, et éventuellement de pointer les décalages, de deux ordres : opérateurs inadaptés au fonctionnement de la langue, et propriétés linguistiques mal ou pas décrites par les opérateurs.

Les opérateurs ne sont pas tous exactement les mêmes d'un moteur de recherche à l'autre. En général, soit ils diffèrent par leur expression (AND et ET) — ce que nous pouvons négliger ici sans dommage —, soit certains opérateurs spécifiques sont ajoutés, en complément des opérateurs les plus connus — nous mentionnerons par exemple des opérateurs du moteur TOPIC (de la société Verity), très complet, pour faire à peu près le tour de l'existant.

### 3.2. Les fonctions de rectification

Le linguiste distingue deux dimensions de la langue : la dimension *syntagmatique*, qui est l'enchaînement linéaire des mots au fil d'un texte ; et la dimension *paradigmatique*, qui correspond aux mots apparentés (par leur sens, leur nature morphosyntaxique) susceptibles de figurer en un point donné du texte.

Une première série d'opérateurs s'applique à corriger les décalages introduits par la non correspondance entre les chaînes de caractères et les mots.

Au plan syntagmatique, le découpage a démantelé toutes les expressions composées qui fonctionnent d'un seul tenant dans la langue, en particulier les locutions ('bien que') ou les figements ('(en) file indienne'). Le rôle de l'opérateur consiste à ne plus prendre en compte que la forme complète et telle quelle, sans plus considérer les composants séparément (par ex., ne pas voir 'bien' dans 'bien que', ni 'indienne' dans 'file indienne').

Au plan paradigmatique, il s'agit de confondre des formes qui ont une graphie différente (une abréviation et sa forme développée, des variantes d'écriture comme 'clé' vs 'clef'), ou au contraire de rétablir un contraste normalement négligé par le processus d'indexation (typiquement l'utilisation des majuscules). Dans le même esprit, certains moteurs plus sophistiqués proposent des opérateurs pour rétablir l'équivalence de variantes explicables par des propriétés de la langue : son écriture alphabétique et sa redondance (rétablir l'identité par delà une altération portant sur un caractère : calcul de distances floues entre chaînes de caractères), sa dimension phonétique (lorsque l'identité du mot est surtout sonore, et est susceptible de retranscriptions diverses : utilisation pour des noms propres étrangers, par exemple).

Ces rectifications (r)établissent des identités entre formes graphiques : c'est une redéfinition complète (le temps de la recherche) des unités d'indexation. Cependant, les interrelations syntagmatiques et paradigmatiques des mots peuvent être plus souples (existence d'une relation mais aussi existence à part entière des composants), ce qui motive une seconde gamme de fonctions.

### 3.3. Les fonctions d'interrelation

Au plan syntagmatique, il existe des formes composées susceptibles de connaître des petites variations d'usage : par exemple, insertion d'un complément, élision ou ellipse d'une composante. Il n'y a pas d'opérateur pleinement approprié à la description de ce comportement linguistique, pour lequel on recourt à des fonctions de distance syntagmatique quand elles sont disponibles (par exemple,



les composants sont séparés par au plus trois mots), ou à l'opérateur de concaténation s'il fonctionne par exemple en ne tenant pas compte des mots grammaticaux ("amour art\*" pour 'amour de l'art', 'amour des arts', 'amour pour l'art', etc.).

Les opérateurs les plus utilisés sont sans doute les opérateurs paradigmatiques qui permettent de saisir l'ensemble des formes fléchies d'un mot : les formes conjuguées d'un verbe, la variation singulier / pluriel pour un nom, à laquelle s'ajoute la variation en genre pour l'adjectif. La troncature (souvent notée par le caractère \*), qui fixe le radical et laisse indéterminée la terminaison, est une approximation, qui a l'avantage d'économiser l'énumération, souvent fastidieuse et pas toujours complète. Lorsque la troncature est appliquée en amont du suffixe, et aussi éventuellement au niveau du préfixe, elle sert à décrire les mots d'une même famille.

Lorsque le moteur le permet, il est recommandé de visualiser l'ensemble des graphies correspondantes dans le corpus, pour éliminer en amont les mots ainsi saisis mais sans rapport avec la recherche, plutôt que d'avoir à les écarter de façon démultipliée en aval, au moment du dépouillement des résultats. L'examen des occurrences correspondants à la forme tronquée renseigne sur l'opportunité d'utiliser cet opérateur et sur l'endroit le plus adéquat où l'appliquer.

Ce passage par l'examen des occurrences vient de ce que la troncature définit un ensemble en compréhension et non en extension : autrement dit, c'est un procédé, une règle de dérivation, qui indique l'équivalence entre deux mots, et non leur mention explicite d'appartenance à une même liste. Il existe aussi des opérateurs qui définissent des paradigmes en extension, typiquement l'opérateur OU (disjonction non exclusive). On utilisera donc le OU pour grouper des mots dont on considère qu'ils jouent le même rôle dans la recherche, qu'ils sont en quelque sorte interchangeables.

Les moteurs peuvent inclure des dictionnaires de synonymes, ou des classes de mots associés déterminés par un calcul statistique. Un opérateur<sup>10</sup> convoque alors, pour le mot auquel il est appliqué, l'ensemble des mots qui lui correspondent. C'est une aide à l'enrichissement et l'extension de la requête, pour rejoindre la diversité des expressions d'une même idée dans un texte.

Notre analyse distingue donc des fonctions d'identité, pour lesquelles la variation graphique est considérée comme absolument non significative, et des fonctions d'équivalence, qui confèrent un rôle analogue aux mots qu'elles regroupent dans le cadre d'une recherche et d'un corpus. Il est important de prendre conscience de cette différence, importante au plan linguistique, alors qu'elle ne trouve guère d'écho dans les opérateurs. En particulier, les relations de synonymie ou de variations flexionnelles (singulier/pluriel, etc.) et dérivationnelles (familles de mots) sont de l'ordre de l'équivalence plutôt que de l'identité : il s'agit de différences que l'on choisit momentanément de négliger, mais qu'il peut s'avérer intéressant de distinguer et de considérer tour à tour au moment de la consultation des contextes.

---

<sup>10</sup> Par exemple, respectivement, les opérateurs SYNONYMES et SUGGESTION, dans le moteur TOPIC de la société Verity.

### 3.4. Les fonctions de contextualisation

Vient enfin un opérateur très connu et dont nous n'avions pas encore parlé. A la réflexion, le ET stipule la cooccurrence dans un même texte : c'est donc une contrainte de contexte.

Les moteurs les plus riches en opérateurs affinent cette contextualisation en déclinant essentiellement deux autres types de contextes. Un contexte de l'ordre du paragraphe, ou de la page, évite des cooccurrences très lâches (de part et d'autre du texte), en situant les occurrences dans un empan typiquement homogène au plan de la thématique. Un contexte de l'ordre de la phrase (ou d'une fenêtre de quelques mots) est utile pour rechercher des dépendances syntaxiques souples.

Ces trois niveaux de contexte (texte, paragraphe, période) sont sémantiquement pertinents, et correspondent à des interrelations de natures différentes. Concrètement, leur définition (et l'automatisation de leur repérage) est un délicat équilibre entre des informations linguistiques (syntaxe, ponctuation), typographiques (alinéa), perceptives et cognitives (rapport entre la taille des contextes, le champ visuel, et les différents types de mémoire : mémoire de travail, mémoire à long terme). Les délimitations concordent avec des seuils (une ponctuation forte, un retour à la ligne) mais ont une certaine perméabilité : par exemple, la dernière phrase du paragraphe précédent peut être en relation contextuelle avec le paragraphe considéré. Le choix des moteurs est en général simplifié, en ne considérant qu'un mode de définition à la fois : par exemple, voisinage dans la même phrase (les phrases étant découpées selon les ponctuations fortes), voisinage dans le même paragraphe (délimitation par retour à la ligne ou saut de ligne), voisinage dans une fenêtre de  $n$  mots.

Le contexte joue un rôle capital au plan sémantique. Lorsque plusieurs mots partagent un même contexte, ils se désambigüisent réciproquement et évitent en grande partie la dispersion des résultats attribuée à la polysémie et à l'homonymie. C'est pourquoi toute recherche thématique, qui ne serait pas uniquement focalisée sur les emplois d'une expression précise, a fortement intérêt à lancer l'exploration à partir de plusieurs mots conjoints.

Les travaux d'Evelyne Bourion présentés ici même il y a deux ans (Bourion 2000) nous en donnent une illustration concrète. Partant du verbe 'parer' sous ses différentes formes conjuguées, un test statistique sélectionne<sup>11</sup>, au voisinage de ce mot dans un corpus de romans, des mots associés, qu'Evelyne Bourion analyse *a posteriori* comme se rapportant à plusieurs domaines : la thématique de la //parure// ('pomponné', 'bijou', 'beauté', 'joyau', 'coquetterie', 'charme',...), mais aussi l'//obstacle// (parer un coup : 'spadassin', 'coup', 'botte', 'choc',...), la //marine// ('canot' pour l'expression 'parer un canot'), voire même l'//automobile// ('pare-brise', et avec lui 'rétroviseur'). Relancer une recherche sur 'parer' accompagné du vocabulaire du domaine visé (ici, celui de la parure) aurait permis d'écarter automatiquement les homonymes non désirés.

C'est hélas à ce stade de la contextualisation que le booléen révèle toutes ses limites. Lorsqu'un ET conjoint plusieurs mots, si *presque*

---

<sup>11</sup> Les contextes du verbe 'parer' sont examinés dans un grand corpus de romans. Le calcul repère les mots qui figurent de façon « anormalement fréquente » (test statistique de l'écart réduit) dans un voisinage de 10 mots de part et d'autre de 'parer', au sein de la même phrase (les voisinages s'arrêtent aux ponctuations fortes).

*tous* ces mots se trouvent dans un texte, c'est exactement comme si *aucun* n'y figurait : le texte n'est pas retenu. Pour une liaison moins contraignante, on dispose du *OU* : mais alors, retrouver *tous* les mots n'est en rien distingué de la présence d'*un seul* mot (hors contexte). Assouplir la requête se paie par une combinatoire fastidieuse et jamais pleinement satisfaisante non plus<sup>12</sup>.

Bref, le *ET* impose la présence, ce qui est inadéquat pour une recherche thématique sur le texte intégral. Symétriquement<sup>13</sup>, l'opérateur *SAUF* ou *NON* a pour rôle d'exclure la présence, ce qui est tout autant dangereux pour la recherche thématique en texte intégral (prétérition, homonymie, mention secondaire, etc.).

Cette inadéquation criante a valu la mise au point de nouveaux opérateurs documentaires, tel l'opérateur *CUMUL* du moteur *TOPIC*. Il s'agit bien de lier les mots sur un mode intermédiaire entre le *OU* et le *ET*, en valorisant leur diversité (ce que ne fait pas le *OU*) sans forcer leur présence (pour éviter les déconvenues du *ET*).<sup>14</sup>

Plus prometteuse encore s'annonce la perspective de l'interrogation par un texte, à l'image de la technique très ancienne des passages parallèles<sup>15</sup>. L'expression du thème de recherche est alors tout bonnement un texte représentatif. Le fonctionnement du moteur consiste à s'appuyer sur les associations contextuelles des mots dans ce texte pour identifier des associations analogues dans le fonds à explorer. La sélection ne se fonde plus sur un principe d'identité et de reconnaissance de motifs prédéfinis, mais elle admet, plus justement, une part d'implicite et de variation, cadrés par le contexte global du vocabulaire. Cette manière d'aborder la recherche en texte intégral évite de se focaliser sur quelques mots fixés, au bénéfice d'une plus grande sensibilité à la sémantique d'ensemble d'un passage ou d'un texte. La qualité des résultats est supérieure à ce que l'on obtient par une interrogation limitée à quelques mots, grâce à cette contextualisation souple : l'expérience positive du moteur *DECID* en est une première illustration (Bommier-Pincemin 1999).

Dans l'état actuel des outils disponibles, on préférera donc le vectoriel non structuré pour une recherche thématique. Le booléen garde ses attraits pour retrouver un certain texte (démarche non exploratoire) à partir d'une caractéristique discriminante, ou encore pour une recherche centrée sur une expression déterminée<sup>16</sup>. Dans le cadre d'un travail littéraire, le booléen s'accorde avec une recherche concernant un signifiant (tel mot, telle expression précise, tel

---

<sup>12</sup> Entre « A ET B ET C ET D » (trop restrictif) et « A OU B OU C OU D » (trop lâche), il y aurait « (A ET B ET C) OU (A ET B ET D) OU (A ET C ET D) OU... »

<sup>13</sup> Il est symptomatique que certains moteurs, comme AltaVista, aient introduit les notations + et -, comme opérateurs s'appliquant à un mot, respectivement pour imposer ou exclure sa présence.

<sup>14</sup> Concrètement, *CUMUL* (parait, bijou, charmes) permet de trouver les textes où figurent tout ou partie des mots indiqués, tout en mettant en valeur les textes comportant les trois mots, puis ceux en comportant deux, puis en dernier ressort en considérant ceux qui n'en ont qu'un.

<sup>15</sup> Face à un passage obscur dans un texte, le lecteur part à la recherche d'autres passages abordant le même sujet, susceptibles de lui apporter d'autres éléments de compréhension.

<sup>16</sup> Le vectoriel et le booléen ne s'opposent finalement pas tant au plan de la qualité des résultats, qui permettrait de conclure à la supériorité de l'un sur l'autre, qu'au plan de leur différence de perspective. D'où le caractère artificiel que prendrait un banc d'essai, visant à évaluer leurs mérites respectifs.

suffixe), le vectoriel convient mieux lorsque c'est un signifié qui est visé (typiquement un thème).

## **4. Exploitation de résultats : aides et dangers**

### **4.1. Ordre et tris**

La présentation des résultats de la recherche « par ordre de pertinence », *i.e.* en commençant par ceux qui correspondent le mieux à la requête, est vantée comme le *nec plus ultra* des moteurs de recherche. Ce classement par score de pertinence décroissant n'est pas sans fasciner par son apparence objective, la technicité et l'expertise de l'appareil numérique affleurant, nimbées d'un précieux mystère (secret commercial oblige !). A l'usage pourtant, cette organisation linéaire des réponses est décevante. Elle distord la nature du jugement de pertinence, qui est multidimensionnel (lors d'une recherche différents critères d'intérêt se conjoignent, qu'il est artificiel de hiérarchiser : tel texte proposé est intéressant pour tel aspect, tel autre pour tel autre aspect complémentaire, etc.). Ceci engendre un parcours inefficace des résultats : la liste mélange et interclasse différents « filons », si bien que toute erreur se répercute de façon diffuse. Considérant chaque texte l'un après l'autre, l'utilisateur ne dispose pas d'une vue d'ensemble, qui lui permettrait d'orienter son dépouillement, et de savoir où s'arrêter.<sup>17</sup>

Les procédures de tris, bien que plus frustes, se révèlent plus intéressantes à exploiter. Les tris selon différents champs se prêtent à des vues diverses et complémentaires des résultats : chronologie, localisation (même auteur dans une collection de textes littéraires, même url sur internet), ou pourquoi pas, comme indicateur de pertinence, informations sur les mots trouvés dans le texte et sur leur contexte.

### **4.2. L'indispensable retour au texte**

Comme nous l'avons déjà souligné, la consultation des textes repérés par le moteur est d'une importance capitale pour l'interprétation des résultats. Ceci a bien été compris, et les interfaces hypertextes sont mises à profit en ce sens : un lien facilite l'accès au texte, via un simple clic souris.

Autre exemple significatif : un des points forts qui font l'attrait du moteur Google<sup>18</sup>, est la présentation, pour chaque page internet proposée, d'un extrait caractéristique comportant les mots de la requête (surlignés) : cette information est extrêmement efficace pour aider l'utilisateur à dépouiller les résultats.

### **4.3. Les représentations graphiques dans le plan (cartographies)**

La présentation visuelle d'un ensemble de textes ou de thèmes sous la forme de leur disposition dans un espace déborde quelque peu la description des moteurs de recherche standard. Mais il nous semble bon de lui consacrer un paragraphe, étant donné le développement actuel de ces cartographies, et l'attrait qu'elles exercent.

---

<sup>17</sup> Ce n'est pas le lieu de développer ici le concept alternatif de « pertinence différentielle » : il consiste à présenter les axes majeurs qui structurent les résultats de la recherche, puis à mettre en valeur les spécificités originales des textes proposés pouvant motiver leur intérêt (Bommier-Pincemin 1999).

<sup>18</sup> <http://www.google.com/>

Ces représentations allient en effet séduction esthétique (couleurs)<sup>19</sup>, évocation imagée (plusieurs métaphores ont été adoptées : îles et continents dans un océan<sup>20</sup>, relief d'un massif montagneux avec ses cols et ses sommets<sup>21</sup>, molécules chimiques ou structures futuristes se détachant d'un espace étoilé<sup>22</sup>), et apparente simplicité d'interprétation à partir de la signification intuitive des proximités, des positionnements centrés ou polaires, des oppositions. Or il faut accorder la plus grande attention aux conventions de représentation, et aux codes de lecture qui doivent guider l'interprétation.

Considérons par exemple les réductions et choix de représentation qui s'imposent successivement pour la représentations de plusieurs points dans un plan.

Avec deux points déjà se pose la question de l'échelle, pour se faire une idée juste du degré de proximité ou d'éloignement.

Avec trois points, apparaissent des considérations proprement mathématiques. Si la mesure qui évalue l'espacement entre deux points n'est pas un *écart*, ni une *distance*, mais seulement un *indice de dissimilarité*, alors l'inégalité triangulaire n'est pas vérifiée ; et pour une série de valeurs décrivant l'espacement entre ces trois points deux à deux, il n'est pas dit qu'on puisse les représenter en respectant toutes ces mesures.

Pour quatre points, dans le cas général ils ne sont pas tous dans un même plan : trois points (non colinéaires) définissent un plan précis, auquel un quatrième point peut appartenir ou non (un trépied n'est jamais bancal — il positionne ses trois points d'appui sur le « plan » du sol —, alors que les chaises et tables à quatre pieds peuvent l'être, dès que leur quatre points d'appui ne sont plus ajustés dans un même plan).

Intervient donc une projection, selon une certaine perspective : tout l'art de calculs comme les analyses factorielles consiste à trouver une représentation qui minimise les inévitables déformations. L'interprétation doit se prémunir contre deux illusions d'optique :

- l'effet étoile double : pour l'observateur ayant une bonne vue (ou une correction satisfaisante) l'avant-dernière étoile de la queue de la grande ourse se dédouble en deux étoiles très proches, Mizar (la plus brillante) et Alcor (en arabe 'l'épreuve' — c'est un test d'acuité visuelle !). Or ces deux étoiles, qui paraissent si proches au point de se confondre pour l'observateur non averti, forment en fait un couple optique : elles sont séparées par des dizaines d'années lumière. L'une est loin derrière l'autre, mais, du fait du point d'observation (la terre) et de la ligne de visée, elles figurent l'une juste à côté de l'autre. C'est exactement le même phénomène qui se produit dans les représentations planes : la proximité de deux points dans le plan, après un calcul de réduction d'un espace multidimensionnel, ne permet pas de conclure à leur proximité effective, il peut rester des dimensions non représentées et qui les écartent. Des aides à l'interprétation

---

<sup>19</sup> Quelques exemples hauts en couleur, appliquant les Cartes Auto Organisantes (Self Organizing Maps) de Kohonen à des collections de documents, sur le site Websom : <http://websom.hut.fi/websom/>

<sup>20</sup> Umap (de la société Trivium, dans la lignée des 'arbres de connaissances') : <http://www.umap.com/>

<sup>21</sup> ThemeScape (<http://www.newsmaps.com/>), appliquant une technique présentée à l'adresse <http://www.cartia.com/>. Il y a même des petits drapeaux, pour repérer un territoire !

<sup>22</sup> Sémiomap génère des graphes « élastiques » dans un espace interstellaire : <http://160.149.176.110/semiomap/discovery.htm>

sont prévues, par exemple visualiser les plus proches voisins d'un point indiqué en les reliant à lui, ou consulter l'indicateur *cosinus carré* de l'analyse factorielle qui renseigne sur la proximité réelle à chacun des axes.

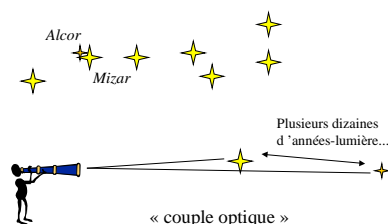


Figure 2 : L'effet étoile double

- l'égalisation des magnitudes : il ne suffit pas de positionner un point, il faut aussi savoir quelle est son importance, et son incidence sur la structure d'ensemble. Pour reprendre l'image de la voûte étoilée, si l'on en faisait une carte en représentant toutes les étoiles par des points semblables, sans faire la part entre les étoiles les plus brillantes et celles qui le sont moins, la carte serait illisible, on n'y verrait plus les motifs des constellations. Dans le contexte d'une représentation spatiale de mots ou de textes, il est important de savoir quels éléments ont pris une importance décisive dans l'articulation de la représentation, quels points sont influents, caractéristiques d'une région de l'espace. Là encore il existe des indicateurs graphiques (taille des points) ou numériques (par exemple les *contributions* pour l'analyse factorielle) qui étayent l'interprétation.

## 5. Synthèse et généralisation : trois « angles d'attaque » du texte intégral

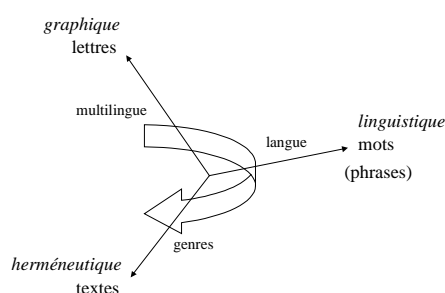


Figure 3 : Trois perspectives sur les écrits numériques

### 5.1. Perspective graphique : les lettres

La perception des textes comme des suites de lettres correspond au point d'accès le plus évident dans les traitements informatisés<sup>23</sup>. Pour

<sup>23</sup> Notre exposé se centre sur le cas des écrits, sans que cela remette en question la pertinence des corpus oraux. Quelle pourrait être la « perspective graphique » dans ce cas ?

L'écrit peut être médiateur, via un travail (considérable) de retranscription. Pour l'écrit lui-même, la représentation alphabétique et le codage des caractères opèrent déjà une perception élaborée du texte : que l'on considère simplement les difficultés dans le domaine de la reconnaissance de l'écriture

le codage informatique, les « données textuelles » sont des chaînes de caractères alphanumériques, de longueur indéfinie.

Evidemment, cette réduction très matérielle et appauvrie de la réalité textuelle prête le flanc à toutes les critiques. Un texte n'est pas un simple alignement de lettres, d'ailleurs la génération aléatoire d'un chapelet de lettres n'aurait guère l'allure d'un texte.

Ceci dit, les techniques frustes ne sont pas sans intérêt quand elles se basent en fait sur des phénomènes linguistiques. Représenter un texte par la suite de ses lettres et de ses mots graphiques (suites de lettres entre blancs ou ponctuations) rend compte de la dimension écrite de la langue. C'est un modèle s'appuyant sur l'écriture alphabétique, et la double articulation<sup>24</sup>.

En particulier, les langues se caractérisent comme un « code » redondant : le contexte permet de rectifier des altérations. Cette robustesse est favorable à la réalisation d'observations significatives, malgré toutes les réductions opérées pour la représentation du texte.

Par exemple, la technique des *n*-grams consiste à représenter le texte comme le décompte des suites de *n* caractères alphabétiques qui y figurent, à l'intérieur et aux extrémités des mots (mais ne relevant pas de plusieurs mots). Cette décomposition, qui donne une image du texte quasiment illisible pour l'œil humain, donne cependant des résultats étonnants dans des systèmes de recherche d'information. Comme typiquement *n* vaut deux ou trois (voire quatre), ces bribes de mots reflètent des syllabes et des racines, et la modélisation rejoint donc certains aspects tout à fait linguistiques.

Ce point de vue graphique fournit des outils et traitements multilingues, ou du moins applicables à des familles de langues. Tel outil est ainsi capable de proposer des analyses de tout texte, pourvu qu'il soit transcrit avec un alphabet. Certaines applications ont même été développées en jouant du caractère interlingue des racines, en considérant de façon unifiée des collections de textes de langues romanes.

## 5.2. Perspective linguistique : les mots

C'est donc un autre point de vue qu'introduit la prise en compte du système lexical et grammatical d'une langue. Les mots sont reconnus en tant qu'entités linguistiques, par delà leur forme graphique dans le texte. Les relations et structurations syntagmatiques et paradigmatiques entrent en ligne de compte. La description de l'organisation interne de la phrase, voire de l'enchaînement successif des phrases, est du même ordre. Le texte est considéré sous l'angle de la langue dans laquelle il est rédigé, avec son vocabulaire ses régularités, ses transformations et ses structures.

---

manuscrite, où la délimitation et l'identification des mots et des lettres est extrêmement complexe.

L'analyse directe de composantes captées et mesurées dans le signal sonore pose la question du choix pertinent des caractéristiques saisies. Pour l'écrit, la représentation alphabétique n'est finalement qu'une « bonne » perception des textes, parmi une infinité d'autres perceptions possibles : la couleur de l'encre, la densité d'occupation des pages... Pour l'oral, les tentatives d'utilisation de courbes sonores pour l'apprentissage des langues étrangères montre par exemple que les paramètres qu'elles mesurent ne correspondent pas à une perception significative, puisque les courbes obtenues pour une même phrase sont extrêmement diverses, et que rien ne permet de distinguer une prononciation correcte ou incorrecte.

<sup>24</sup> En simplifiant, le texte s'articule en mots, et chaque mot se décompose en lettres.

Dans le cas des moteurs de recherche, l'intervention d'une telle perception linguistique<sup>25</sup> des textes prend la forme de modules de :

- catégorisation : identification des mots et de leur nature, dans le but notamment de séparer les homographies nom / verbe, préposition / nom, etc.
- lemmatisation : homologation des variantes flexionnelles (en genre, nombre, personne, temps)
- repérage des syntagmes, dans l'optique de travailler également sur des termes dont la signification ne se réduit pas à la composition des mots dont ils sont formés, et de gagner en précision sémantique en prenant en compte des interrelations étroites entre les mots.

Les recherches et les réalisations dans le domaine du Traitement Automatique des Langues ont essentiellement porté sur les descriptions lexicales, morphologiques et syntaxiques. Les traitements s'effectuent mot à mot (en isolant des chaînes de caractères et en consultant un dictionnaire, éventuellement un analyseur morphologique), ou phrase à phrase (le texte est analysé « par morceaux », en itérant la même procédure d'analyse ; le produit du traitement sur le texte est le cumul des résultats obtenus successivement).

Or les mots ne se construisent pas (uniquement) à partir des lettres, mais à partir de l'environnement textuel. Cela joue aussi bien au plan du signifiant, pour leur délimitation ([Bayard] [monte] au [créneau] vs [Rocard] [monte au créneau], cf. Rastier 1997), qu'au plan du signifié, pour la constitution de leur sens, le repérage de leur acception (c'est l'objectif de l'exercice de construction des champs sémantiques d'un mot à travers une œuvre, une collection de textes). L'environnement contextuel global module aussi le relief que prend à la lecture tel ou tel mot.

### 5.3. Perspective herméneutique : les textes

Si l'on prend acte que le texte ne se définit pas uniquement comme une chaîne de caractères, ni comme une « longue » suite de mots ou de phrases, s'ouvre alors une troisième perspective, proprement textuelle.

Ainsi, chaque texte se présente pour le lecteur comme une entité cohérente, un tout. En ce sens, un corpus n'est pas conçu comme *du* texte (« au kilomètre »), mais comme *des* textes. Le texte se voit reconnaître une certaine autonomie, qui contribue d'ailleurs à justifier les calculs et traitements sur corpus. Dans cet esprit, une approche selon la sémantique différentielle (Rastier & al. 1994) permet une sémantique pleinement linguistique, non fondée sur une description ontologique du monde et de l'environnement situationnel. Ce n'est pas que le texte soit sans relation avec « l'extérieur » : les différents pôles non linguistiques utilisés pour le caractériser (l'auteur, les lecteurs, le monde) se reflètent comme en creux dans l'expression linguistique du texte, comme autant de « pôles intrinsèques » (Rastier 1996), notamment via la forme d'un genre. Aussi la perspective textuelle ne suppose-t-elle souvent plus la langue homogène et égale, mais élabore des traitements dans le cadre des genres textuels.

---

<sup>25</sup> Le qualificatif « linguistique » a été choisi ici pour son caractère évocateur, mais est utilisé selon une acception restreinte du terme. L'analyse textuelle (décrite dans la troisième perspective) entre aussi dans le champ de la linguistique.



La délimitation externe du texte est une articulation centrale d'un jeu de paliers de contextes, qui comprend aussi des zones plus étendues (l'intertexte) ou plus étroites (le paragraphe). Ces voisinages ont une incidence sémantique en structurant les interrelations entre les mots. La perspective textuelle est donc aussi celle de la prise en compte des effets de contexte, de la contextualisation des mots.

Cette perspective souligne tout à la fois l'importance et la relativité du corpus. Autrement dit, le corpus, en tant que contexte intertextuel « opérationnel », est déterminant pour l'analyse des textes. Mais aussi, il ne reflète jamais qu'un point de vue, qu'un contexte d'analyse parmi d'autres possibles, et aux déterminations différentes. La perspective textuelle introduit une conception herméneutique du texte, dans laquelle la lecture est un acte de parcours interprétatif et de construction d'un sens, ni exclusif, ni arbitraire.

A cette perspective textuelle se rapporte la convergence d'intérêt pour l'étude et l'analyse des contextes, pour différents outils présentés dans cette journée. Un rôle central et médiateur est donné à l'utilisation des contextes des mots dans les traitements, ou à l'observation des contextes pour l'exploitation des résultats des calculs.

Les concordances sont un outil traditionnel et essentiel : des logiciels se spécialisent dans leur construction et leur maniement (X-COR), mais les autres logiciels travaillant sur les textes ne sauraient s'en passer (INTEX comme Hyperbase, aussi différents soient-ils par ailleurs, établissent des concordances). L'objectif premier des concordances est de focaliser l'attention sur les occurrences d'une unité linguistique en contexte. La présentation (centrée sur le mot dont on examine les contextes, superposant les lignes, et triée) fait ressortir les parallélismes et les récurrences.

Une des fonctionnalités essentielles d'Hyperbase, activée par la commande « thèmes », est le calcul statistique des corrélats sémantiques d'une unité. Là encore, la pertinence de ce traitement repose sur le rôle sémantique des contextes.

Et en ce qui concerne les moteurs de recherche, le présent exposé a expliqué l'intérêt de rechercher des groupements souples de mots, avec des opérateurs de voisinage et un opérateur intermédiaire entre le OU et le ET comme l'opérateur CUMUL. Les mots pris isolément ne donnent pas un accès sémantique satisfaisant. La formulation d'une recherche thématique ne repose pas sur l'expression de quelques mots, mais sur l'évocation d'un contexte. En particulier, un thème peut être évoqué dans les textes de façon synthétique ou diffuse, complète ou partielle : il n'est pas nécessairement lexicalisé, voire même sa lexicalisation est évitée lorsque l'auteur fuit les formulations triviales, et que l'effet littéraire vise l'évocation plutôt que la désignation.

#### **5.4. Rapports entre ces trois perspectives**

Comme voudrait le représenter le schéma, les trois perspectives (graphique, linguistique, herméneutique) sont complémentaires. Trop souvent la perspective linguistique occulte les deux autres : il convient de réhabiliter un usage judicieux de la perspective graphique, comme de faire percevoir l'apport de la perspective textuelle, encore méconnue.

## Références bibliographiques

- BOMMIER-PINCEMIN Bénédicte (1999) - *Diffusion ciblée automatique d'informations : conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*, Thèse de Doctorat, Linguistique, Université Paris IV (Sorbonne), 6 avril 1999, n°99PA040027, 806 pages.
- BOURION Evelyne (1998) – « Ponctuation et accès sémantique aux banques textuelles », Actes du colloque *A qui appartient la ponctuation ?*, Liège, 12-14 mars 1997, Paris, Bruxelles, Duculot, pp. 409-435.
- BOURION Evelyne (2000) – « Corpus électroniques et lecture non-linéaire : vers une assistance à la recherche thématique », Actes de la Journée Scientifique « Philologie électronique et assistance à l'interprétation des textes », *Recherches en Linguistique et Psychologie cognitive*, 15, Presses Universitaires de Reims, 198-223.
- MULLER Charles (1985) - Allocution inaugurale, Actes du Colloque *Méthodes quantitatives et informatiques dans l'étude des textes – Hommage à Charles Muller*, Nice, 5-8 juin 1985, Genève - Paris : Slatkine - Champion, 7-12.
- PINCEMIN Bénédicte (1999) - « Construire et utiliser un corpus : le point de vue d'une sémantique textuelle interprétative », Actes de l'Atelier *Corpus et TAL : pour une réflexion méthodologique*, A. Condamines, M.-P. Péry-Woodley et C. Fabre (éd.), TALN'99, Cargèse, 12-17 juillet 1999, 26-36.
- PINCEMIN Bénédicte, LEMESLE Xavier (1996) - *Etat de l'art des idées implémentées dans les moteurs de recherche par index sur WWW*, Collection de notes internes de la Direction des Etudes et Recherches, 97NO00011, EDF Clamart, 37 pages, ISSN 1161-0603.
- RASTIER François (1987) - *Sémantique interprétative*, Paris : Presses Universitaires de France, 277pages.
- RASTIER François, CAVAZZA Marc, ABEILLE Anne (1994) - *Sémantique pour l'analyse – De la linguistique à l'informatique*, Paris : Masson, coll. Sciences cognitives.
- RASTIER François (1996) - « Pour une sémantique des textes - questions d'épistémologie », in *Textes & Sens*, François RASTIER (dir.), Paris : Didier Erudition, 9-35.
- RASTIER François (1997) - « Défigements sémantiques en contexte », in Martins-Baltar M., éd., *La locution entre langue et usages*, coll. Signes, E.N.S. Fontenay-St Cloud Editions, diff. Ophrys, 305-329.
- SALTON Gerard, MCGILL Michael J. (1983) - *Introduction to Modern Information Retrieval*, New-York : McGraw-Hill.